A Review of Data Science in Business and Industry and a Future View

Grazia Vicario[1,*,+] Shirley Coleman[2]

[1] Department of Mathematical Sciences, Politecnico di Torino. C.so Duca degli Abruzzi 24, 10129 Torino (Italy)
[2] School of Mathematics, Statistics and Physics, Newcastle University, NE1 7RU, UK

Abstract

The aim of this paper is to frame Data Science, a fashion and emerging topic nowadays in the context of business and industry. We open with a discussion about the origin of Data Science, and its requirement for a challenging mix of capability in data analytics, information technology and business know-how. The mission of Data Science is to provide new or revised computational theory able to extract useful information from the massive volumes of data collected at an accelerating pace. In fact, besides the traditional measurements, digital data obtained from images, text, audio, sensors, etc complement the survey. Then we review the different and most popular methodologies amongst the practitioners of Data Science research and applications. And since the emerging field requires personnel with new competences, we attempt to describe the Data Scientist profile, one of the *sexiest jobs of the 21st Century* according to Davenport and Patil. Most people are aware of the need to embrace Data Science, but they feel intimidated that they don't understand it and they worry that their jobs will disappear. We want to encourage them: Data Science is more likely to add value to jobs and enrich the lives of working people by helping them make better, more informed business decisions. We conclude the paper by presenting examples of Data Science in action in business and industry, to demonstrate the collection of specialist skills that must come together for this new science to be effective.

Keywords: Knowledge Discovery, Data Scientist profile, Business improvement, Industry 4.0, SME.

1. Introduction and Rationale of a Data Science review

The paper has been motivated by the need to frame the research area of Data Science in the context of business and industry. In the last three decades, Data Science has grown up and expanded its research area attracting interests and researchers from many neighboring scientific fields. This increasing interest in Data Science arises both from public and private organizations. For example, an internet shop can exhibit custom-made products and focused advertisements if it is possible to interpret the

[*] Correspondence to: G. Vicario, Politecnico di Torino, Department of Mathematical Sciences, Corso Duca degli Abruzzi, 24-10129 Turin, Italy.
[+] grazia.vicario@polito.it

data of the customers' web surfing. Or it can forecast demand and improve its logistics management if the sold-out data are properly analyzed.

In the area of healthcare, where therapies and diagnoses are more and more digitized and registered, the use of Data Science methodologies can prevent faulty diagnoses, better detect what are the most appropriate care plans for the patients, and improve the quality of treatments. Also, in industrial production, data coming from different work phases are of primary importance to improve product quality and raise awareness of failures, speed and performance.

Nowadays, the decision processes of organizations are increasingly data-driven. Sensor and processing tools are accessible to small and medium enterprises (SMEs) thanks to the availability of open source software. Big Data support is of fundamental importance to face the competition. The future offers rich possibilities for ever increasing ways of understanding reality, whether it be through increased integration of linked data sets as in Industry 4.0 (1) and the new government drives to link administrative data sets, or through exploring hitherto uncharted data sets (2).

A US survey carried out by KPMG[2] and based on a sample of 400 CEOs highlighted that approximately 77% of them harbored mistrust about the quality of the data upon which their decisions are based. If this were true, all the efforts in analyzing the data would be useless. One problem may be poor communication between those who present the data and those who read about it. People belonging to the data analysis team rarely discuss with the decision makers and this can create an interruption in the communication and development chain. The company realizes the full benefit of Data Science only if the Data Scientists flank the Managers alongside the decision-making process, helping them to understand how the data have been processed. On the other hand, it is also worthwhile for Managers to flank the Data Scientists to reflect the business context in order to optimize the analysis by having shared goals.

In order to help researchers who want to approach the themes and challenges of Data Science and to help people who want to pursue a career as a Data Scientist, we discuss in Section 2 the origin and development of Data Science. In Sections 3, 4 and 5 we review the methodologies, personalities and applications of Data Science, respectively. We provide some examples of Data Science projects in Section 6. Final comments and a future view are given in Section 7.

2. Data Science: origin and development

The continuous data flow produced by the internet and by any sensor attached to modern equipment delivers a huge amount of data. If companies are not able to manage and process the data to their advantage, they will be outperformed by competitors that are. This is the scenario that gave rise to *Data Science* and still supports its development which has been as rapid as the explosive technological change. Data Science is a science offering methodologies for processing and interpreting massive volumes of data collected by an increasing number of new devices. These analytical tasks are difficult to accomplish using just the long-established statistical methodologies. This science is defined today as an interdisciplinary field, including mathematical methods, statistics, algorithm developments, qualitative analysis, computer science and, not less importantly, a practical approach, intending to extract useful information from data, either structured in terms of quantitative information in a set format or unstructured such as reports, visuals and sounds.

Originally, the denomination Data Science was used by the Turing Award's winner Peter Naur in 1960, as a synonym of computer science. Later in 1974, Naur used the term Data Science for

[2] KPMG is a professional service company and one of the Big Four auditors, along with Deloitte, Ernst & Young (EY), and PricewaterhouseCoopers (PwC). From https://en.wikipedia.org/wiki/KPMG.

referring to data processing methods in their full range of applications (3). But it was in 1996 that the term appeared in a public writing for the very first time: the International Federation of Classification Societies organized a meeting in Kobe (Japan) for their biennial conference and named it "Data Science, Classification and Related Methods".

Since then, the international community has adopted the name Data Science to indicate an interdisciplinary field. Nevertheless, Data Science has several times been the focus of different debates with the purpose of defining its distinction (or its standardization) with respect to Statistics (4). Just to mention one instance, C. F. Jeff Wu, during his inaugural lecture for the H. C. Carver Professorship in Statistics at the University of Michigan in 1997 (5), claimed that Statistics should be renamed Data Science and Statisticians Data Scientists. The new modern methodologies, however, are pooling the two disciplines of statistics and computer science as in the interaction of computational algorithms with cognitive science in artificial intelligence and the viewpoint of machine learning as a marriage of statistics and knowledge representation (6).

Leo Breiman, Statistician at the University of Berkeley, has been of a different opinion. In 2001 he said:

> *There are two cultures in the use of statistical modelling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modelling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modelling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.* (7)

Breiman's attitude is reminiscent of a famous phrase by George Box who said:

> *All models are wrong, but some are useful*. (8).

This distinction between Statistics and Data Science is now accepted worldwide, also in university curricula, where classical statistical analysis is taught in courses based on Statistics and new courses on Data Science aim to give a more widely addressed education including algorithms and machine learning.

William S. Cleveland introduced Data Science as an independent discipline in 2001 (9). He considered that six technical areas encompass the field of Data Science: multidisciplinary investigations, models and methods for data, computing with data, pedagogy, tool evaluation, and theory. Subsequently in 2002, the International Council for Science: Committee on Data for Science and Technology (CODATA) founded the first journal on Data Science (https://datascience.codata.org/). Immediately thereafter, Columbia University started the publication of the Journal of Data Science (http://www.jds-online.com/), whose scope is *providing a platform for all data workers to present their view and exchange their ideas*".

Therefore, recently a literature on Data Science has developed, contributing to the information exchange and creating communities joined by the fascination for data, data analysis and its applications. Just to cite one for all, the finance and financial technology community adopts a modern terminology: Fintech Data Science. In fact, one of its representatives, Giudici (10), puts forward the idea of using the terminology Data ScienceS, because Data Science is an integrated process of activities (definition of the objectives of the analysis, selection and processing of the data to be analyzed, statistical modelling and interpretation, implementation and evaluation of the obtained statistical measures), and the different knowledge domains possibly have different objectives expressed in different languages.

Moreover, since the beginning of the 21st century, Data Science has grown in different sub-subjects; among the most important we can mention Knowledge Discovery in Databases, Data Mining, Artificial Intelligence, Machine Learning and Deep Learning, setting the tone for the development of other related areas. In Europe, the European Association for Data Science was founded in 2013, playing an important role in spreading and making more and more popular the use of the terms Data Scientists and Data Analyst, in Europe and especially in the business field.

Summarizing these concepts and opinions, the aim of Data Science is to clean, prepare and analyze different data sets extracting meaning from data, thus being related closely to Statistics; nevertheless, Statistics is used as an instrument by Data Science to reach its purpose, flanked by information technology, in particular programming for big data analysis, and using a distinctly practical approach and a decision making implementation.

3. Data Science: Knowledge Discovery in Data Bases

Since the early 40's, a number of new terms have come into common use, thanks to Information Technology. Thus, we need to distinguish between data (basic element, usually made of symbols), information (outcome of processing data aimed at organizing, interpreting and contextualizing data for acquiring meaning) and knowledge (set of organized and processed information for the purpose of spreading experience, understanding and competences related to practical problems in industry and business). We could imagine a sort of chain of acquaintance that moves from data to knowledge, with the final objective of deciding upon actions, behaviors and decisions. Therefore, the transition from data to knowledge is strategically important and this is the *Data Science* proposition.

To fulfil this objective, Data Science takes advantage of methodologies that are part of disciplines such as Mathematics, Statistics and Computer Science for analyzing and mining data to achieve knowledge. Its impact is so decisive that it is considered as the *fourth paradigm of science*, namely besides science being empirical, theoretical and computational it is also *data-driven*. That's why Data Science may be considered as a member of a larger family of research fields, involving both academic researchers and private companies, namely Knowledge Discovery in Databases (KDD). KDD refers to the process that leads to extracting high-level knowledge from low-level data. Nowadays, it is one of the most attractive, fascinating and ever-evolving fields to work in, thanks to continuous improvement of the data warehouse and ever-growing use of *Big Data* [3] and consequently of Big Data Analytics (aiming at transforming the huge, heterogeneous and persistent amount of data into usable information).

Big Data originates from ever more innovative multimedia and is characterised by being so big that traditional statistical methodologies are no longer usable. Maybe it is for this reason that a group of enthusiasts of Big Data claims that such abundancy of data means that its analysis does not require a theoretical basis. In 2008, Chris Andersen (11), published a provocative paper entitled: *The End of Theory: the Data Deluge that Makes the Scientific Method Obsolete*. He states:

> "***All models are wrong****, but some are useful". So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behavior, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies like Google, which have grown up in an era of massively abundant data, don't have to*

[3] The name *Big Data* refers to huge amounts of data that require ad hoc procedures to be captured and analyzed. In 2013, the term *Big Data* entered the prestigious Oxford English Dictionary, going against the editing rules that a new word needs to be in use for at least 10 years.

> *settle for wrong models. Indeed, they don't have to settle for models at all. ..... Scientists are trained to recognize that correlation is not causation, that no conclusions should be drawn simply on the basis of correlation between X and Y (it could just be a coincidence). Instead, you must understand the underlying mechanisms that connect the two. Once you have a model, you can connect the data sets with confidence. Data without a model is just noise. But faced with massive data, this approach to science — hypothesize, model, test — is becoming obsolete.*

This thinking started off a vibrant debate that is still alive today in the scientific community. In this paper, we do not want to take part actively in this debate either in favour or against the Anderson mindset but we cannot miss the opportunity to express our opinion concerning statistical models. The useful model cited by Box is linked to a physical or simulated experiment with the aim of prediction; it is not required to reproduce the complexity of the phenomenon, but only to give a significant support to the prediction algorithm. Our conviction is that there is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data as stated by Usama Fayyad et al. (12) and other Data Science practitioners.

Most of the work is still handled by humans, who need to make the correct decisions so as not to spoil the data and, most importantly, to achieve an adequate data significance. Appropriate a-priori knowledge of the problem is needed to correctly handle data research, data preparation, selection and cleaning, to be sure not to use inappropriate data, as well as to properly interpret the results of mining, in order not to misunderstand the information extracted by the data analysis. The KDD process in data bases is essential, even if it is very complex. It may require several iterations but it does lead to success in data research problems if human participation and interaction are effective. According to Usama Fayyad et al. who outlines the basic and practical steps of the KDD process suggested in (13), the steps are:

1. Understand the application domain, gain an adequate a-priori knowledge and identify correctly the goal of the KDD process from the customer’s viewpoint;

2. Select or create a target data set on which the discovery is to be performed;

3. Clean and pre-process data: remove possible noise, by collecting enough information about how to identify it; decide strategies about missing data or missing data fields; account for time-sequence information and known changes;

4. Find useful features to represent the data depending on the objective of the task, trying to reduce dimensions or find invariant representations for the data to lighten the research;

5. Find a data mining method based on statistics that conforms to the task of the Step 1 (summarization, classification, regression, clustering, etc.);

6. Perform an exploratory analysis, a model and hypothesis selection, choosing the algorithm(s) and selecting method(s) for searching for data patterns;

7. Do real analysis using data mining methods, (including classification rules or trees, regression, and clustering);

8. Explain the mind pattern, maybe including visualization of data to extend some useful visual information;

9. Act on the discovered knowledge: use it directly, incorporate it into another system for further action, document and report it. Check for and resolve potential conflicts with previous beliefs.

After concluding the initial steps, in which the necessary data has been found, the Data Mining process is the central core applicable to different purposes, for example verification of hypotheses made by the user, discovery of new patterns never seen before or description, if patterns have to be represented in a human-understandable form. Most Data Mining methods are based on tried and tested techniques belonging to Machine Learning, Pattern Recognition or Statistics.

The Data Mining denomination is frequently associated with Data Science, sometimes even confused with it, because of the large amounts of mathematical and computer science methods used by all of these disciplines. However, they are quite different, because the former is a process for discovering patterns in large data sets (with previous knowledge on it), extracting information and transforming it into an understandable structure for further use, typical for appropriate decisions; whereas in contrast, Data Science is a larger family inclusive, for example, of data cleansing, preparation and final analysis on data sets of every type.

Finally, we consider that the mainstream of Data Science is Data-Driven Decision Making, namely the objective to take decisions based on accurate and thorough data analysis rather than on guesswork and intuition.

4. The Data Scientist

After having framed the concept of Data Science, the question arises: who is or, what is expected to be the Data Scientist? We try to provide a Data Scientist profile. The Data Scientist is the person who understands how real problems relate to available data and so can make the best use of data for creating added value. The Data Scientist manages data, guarantees its integrity and accessibility and mines useful information from data to provide knowledge and prediction, and support decision making.

And what are the competences that a Data Scientist is expected to have? First of all, the Data Scientist has to be able to analyze problems in a logical way and with understanding of the real underlying business or practical issues. They have to be able to deal with a large number of techniques. If you Google “data scientist job description”, you can find many web sites listing these skills (about 144,000,000 results (0.48 seconds): Google informs us) and other links for searches related to the Data Scientist job description. The general agreement is that the following statistical methods should be available to everyone considered to be a Data Scientist:

- High dimensional geometry - nowadays, vectors with many components are habitually used for profitably representing situations in data, even if the vector representation is not the natural original choice in collecting the data.
- Singular Value matrix Decomposition, Principal Component Analysis and of course Matrix Algebra should be familiar tools.
- Modeling data with a mixture of Gaussians, a standard tool in Statistics; clustering that refers to partitioning a set of elements into subsets according to a criterion or just identifying existing natural clusters.
- Random Graphs (they are commonly used in many of today’s contexts, such as World Wide Web, Internet, social networks, journal citations,…), Random Trees and Random Walks on Directed Graphs, Markov Chains and Neural Networks.

In addition, competence in *Machine Learning* (ML) is surely desirable and advantageous for the Data Scientist. ML is an interdisciplinary field that unifies applied statistics and computer science, aiming to estimate complicated functions by using some algorithms that can *learn* from previous data. Thus, ML algorithms are able to *adapt* to the problem they are analyzing. They are used for tasks that are too difficult to be solved with fixed programs: the algorithm is not created to do something, but it is created to learn to do something. They have to process vectors with large amounts of data, both in the *training period* and in the *test period*, concepts that come with Big Data. Analytical models are usually developed on a training data set, tested on a test data set and verified on a further data set before being applied to the whole data set to ensure applicability to the full range of data. This concept is less common in traditional statistical analysis where data are often generated for specific cases and may be in short supply, and models are more likely to be assessed by using leave-one-out techniques.

Algorithm accuracy is measured by the *Task* the algorithm is created for (different types of tasks are: *Classification, Classification with missing inputs, Regression, Transcription, Anomaly Detection, De-noising***)**, by the *Performance* the algorithm has on that task (it can vary from case to case, for example is it better to have an output with few important mistakes or one with several medium mistakes?) and by the *Experience* it is allowed to have during the learning process (quantity of training sets the algorithm is allowed to analyze before starting the test set or the real analysis). Of course, the less data needed to get an adequate analysis, the better the algorithm is considered to be.

Just to pinpoint the required competences of the Data Scientist, we mention the distinction between unsupervised and supervised learning algorithms. Unsupervised learning algorithms experience a data set containing features and try to learn useful properties of the structure of the data set, such as the probability distribution from which the data set was generated or a method to de-noise the whole data set. Basically, the algorithm experiences several examples of vector $\boldsymbol{x}$ and attempts to learn the associated probability distribution or some properties of it. Supervised learning algorithms experience a data set containing features in which each example of a random vector $\mathbf{X}$ is associated with a label or a target of an associated value or vector $\mathbf{Y}$; then the algorithm tries to predict $\mathbf{Y}$ from $\mathbf{X}$, usually using the underlying probability distribution. All in all, we have to admit that there is no neat distinction between unsupervised and supervised learning algorithms, because in theory supervised algorithms ought to be guided by the label, while unsupervised should not; but the same algorithms can often be used in both cases, with minimal differences.

There are some problems, especially with the Artificial Intelligence tasks, that ML cannot solve, for various reasons. *Deep Learning* (DL), or *Deep Neural Networks*, is a new autonomous discipline created to overcome the limits of classical ML algorithms; DL derives from ML, but it may be considered part of ML itself. Nowadays, there are a lot of different types of DL algorithms, developed to complete specific tasks, and that obviously adapt to the one they have been created for (object recognition, speech recognition, bioinformatics, and others). DL algorithms may have different architectures, but they refer to training many-layered networks of nonlinear computational units such as Neural Networks do (*input layers*, *output layers* or *hidden layers*, *deep feedforward neural networks*, …. are terms that DL and Neural Networks share). There have been many outstanding successes in DL applications. For example, the ImageNet challenge involves recognizing 15 million images; network analysis with over 150 layers succeeded in making fewer than 5% errors in 2016 thereby beating performance achieved by human beings (14).

For a comprehensive examination of how Statistics has evolved in the last decades, we suggest seeing the recent book of Efron and Hastie (15). The authors discuss the most influential and popular topics in Statistics and how they have been reshaped by modern computing[4]. It is evident that the Data Scientist is not expected to be a Computer Scientist, but has to have a fair degree of familiarity with Information Technology. Summarizing, the Data Scientist does have cross competences in Statistics, Mathematics and Computer Science. Since we live in a digital epoch where anything may be a source of digital data, the professional who is able to manage data and to

[4] Bradley Efron was awarded the International Prize in Statistics and he will receive it during the next ISI (International Statistical Institute) World Statistics Congress to be held in a Kuala Lumpur, August 2019 (http://www.isi2019.org/). He is a professor of Statistics and Biomedical Data Science at Stanford University. The recognition is mainly due to the "bootstrap" method he developed in 1977 for assessing the uncertainty in scientific research. The method had an extraordinary impact across many experimental scientific fields.

extract value from it will be the most coveted in the working environment. Not by chance did Davenport and Patil (16) claim:

> *The Data Scientist: the sexiest Job of the 21st Century.*

The demand for Data Scientists has led Universities to rush to offer new curricula aimed at training Data Scientists. Most of the courses offered are at post graduate level, although undergraduate courses are also emerging. They vary in emphasis and content but the most common, relevant features are solid theoretical foundations in Statistics and Computer Science combined with practical experience of diverse applications. There must be a good two-way exchange between academia and applications whether interdisciplinary within the university or with business and industrial partners, (17). Universities are ideally placed for promoting a valuable interaction between business and industry by including the supervision of postgraduate projects and placements as part of Data Science training.

Data Science Master's degrees are of interest to all University Faculties because of the potential of attracting students and for collaborative projects. To meet the growing interest from students and the need from employers, many courses have been fast-tracked into existence. A nice method of researching the necessary content, need and practicalities was described by Brown (18) at the University of Canterbury, New Zealand. Members of all faculties were invited to a 2-hour session and asked to write down what they considered to be the characteristics of a Data Scientist who successfully completed their prospective Master's degree in Applied Data Science. They then wrote down the characteristics of the likely applicants to their Master's degree and worked out what modules and activities were needed to get from one to the other. The result was a 12 month taught degree with modules including data analytics and digital humanities, and an industry project. The focus was on flexibility and on creativity and it has seen a burgeoning number of applicants and successful graduates.

We conclude reporting a statement of James Stephen Marron, Professor at UNC-CH Department of Biostatistics, University of North Carolina:

> *I think it's time for Data Science to consider the concept of team work. Data Science problems with one person working alone are mostly solved, but for solving big challenges it is necessary to have the work of a team of data scientists with different skillsets.*

It is certainly true that the growing importance of Data Science and its increasing application in many areas, including SMEs, does mean that it touches upon many different fields of expertise. Where data science is firmly embedded, such as in some government services in the UK, there are sub-divisions within the Data Science community containing specialist experts in one or more of the main subject areas. The requirement of being a T shaped person, who has wide knowledge base as well as deep skills, is common in many work areas and is also an asset in everyday life; but in the Data Science context, we consider it to be essential.

5. Development of Data Science in Business and Industry

The development of Data Science has been motivated by the explosion of data in the digital age. It is worth noting that many of the mathematical, statistical and machine learning techniques in Data Science have been around for many years. What has changed is the availability of massive amounts of data that are now stored rather than being merely observed and then overwritten, and more recently the realization that profound insights and business advantage can be gained from analyzing this data.

Data Science as a profession is increasingly being seen as high paid, glamorous and much sought after. This changing tide is encapsulated in the article in the New Yorker shown in Figure 1.

>> Insert here Figure 1>>

Figure 1 The changing tide in professions

Previously it was the creative "Madison Avenue" people who contrived to sell in the most effective way but now behavioral profiling and customer segmentation techniques employed by mathematicians and statisticians are proving that scientifically targeted advertising is much more profitable than even the most elegant and subtle advertising campaigns. New digital displays and fast interactive processing of customer profiles enable promotions to be presented to people just when they are likely to be most susceptible. Hence the math men are overthrowing the mad men.

Data Scientists are now seen as useful and desirable. There has been a lively debate about the relationship between Statistics and Data Science (19). Statistics plays a key role in Data Science and it is worthwhile considering how statisticians should manage that role. Many data analytical practices revolve around opaque solutions in which data is fed into a black box, calculations are made and an answer comes out. Black box techniques mask the algorithms being used and just report an answer. It is true that there can be good reasons for encasing the methodology in a black box, for example if it is highly complex or so that it cannot be interfered with. However, the downside is that black boxes alienate statisticians and encourage a cavalier attitude to statistical detail and theoretical niceties that is frustrating to statisticians and maybe damaging in the long term.

Black box data analytical solutions for prediction lack robustness against changes in influential variables within the data sets. The solutions are often based on assemblies of models and predictions are averaged out using a range of methods, not all of which are appropriate for the type of data being input. It is not easy to find out which predictors have had the major influence on the prediction. One predictor may be much easier to collect than others and have higher quality, but the advantage of using this variable may not be realized unless further checks are made on the solutions to see the effect of swapping variables. Black box users are not forced to check their data before analysis and there is often little emphasis on residual analysis so that variables with gross outliers are not detected. Users may miss data errors and opportunities for finding key subsets and obvious explanations for apparent patterns. But the black box user tends not to mind any of this provided their prediction is better than the one before and seems to work in the short term. In summary, the problems with dependence on black boxes are that

- Algorithms, tuning parameters, subtle effects and assemblies are not accessible;
- Black boxes may be adequate for short term solutions but robustness to change is not certain;
- Skills to understand the analytical methods are not valued and developed.

Business people tend to like black box approaches as they are superficially easier to understand. Statisticians need to reclaim the field rather than let core black box services take over; they need to stand up for the importance of checking models and not just believing the solutions offered. Statisticians need to proclaim the many important roles for which they are needed, such as: to check quality of input data, evaluate costs of obtaining variables and using proxies, conduct sensitivity experiments, construct and validate models.

Companies have a love-hate relationship with Data Science based on primordial fear of numbers and a strong desire for the benefits that analysis can bring (20). We need to encourage staff by emphasizing that there are many positive outcomes from Data Science, such as less waste of time and effort, more time to devote to their main core work and greater profitability and security in employment.

The growth of Data Science affects academia, business and industry; it has a positive influence because it focuses attention on statistical techniques and raises awareness of data in all walks of life in the same way as did Six Sigma, Total Quality Management and other business improvement initiatives (21). Data Scientists need to be numerate and the growth of Data Science encourages enthusiasm for education in all STEM (Science, Technology, Engineering, Mathematics) subjects at all levels. As well as undergraduate and postgraduate degrees, there are also increasing opportunities for vocational training. The growth in Data Science therefore gives a welcome boost to interest in Mathematics and Statistics from high school to tertiary education and in continuing professional development. An example of government interest in this area is the Knowledge Transfer Partnerships funded by Innovate UK. These provide up to 2/3 of the costs of one to three year projects in which a postgraduate research associate is embedded in a company to develop new expertise, supervised by a university academic. These projects are an effective way for SMEs to grow their business with state-of-the-art guidance but less risk and financial outlay (22).

In some senses, we are in a golden era with great opportunities afforded by the expansion of Data Science before data analytics becomes so entrenched that there are fewer creative opportunities for developing bespoke solutions. We need to ensure that the quality and purity of Statistics is maintained not only for our own professional sensibilities but more importantly because it is only when Statistics is applied sensitively and correctly that the optimum benefits of Data Science are achieved.

6. Examples of Data Science in Action

To demonstrate the collection of specialist skills that must come together for Data Science to be effective, we consider some examples. In each case study the business need is clearly stated, the data integration is explained, and then the statistical analysis and the outcomes are described.

Automotive retail sector

Automotive after sales is a business that generates a massive quantity of data. Each day, vehicle owners and service engineers consult catalogues to find the right parts to carry out repairs to their vehicles. In this case study the business partner is a small to medium enterprise (SME) handling big data from catalogue look-ups and other data associated with the automotive after sales market sector. The business motivation is to increase their service offering to customers focusing on some issues that were identified by the customers themselves and on other opportunities that became apparent after applying exploratory statistical analysis and data visualization of the data. Empirical data analysis can be used to explore many different scenarios. Applying IT skills to amalgamate diverse sources of information followed by statistical analysis to identify patterns can bring valuable business insight. In this way, the Data Scientist is helping to improve the SME's current business and also potentially identifying new products and revenue streams (23). We consider three specific examples relating to: mileage; return rates and original equipment manufacturing.

Firstly, looking at the mileage of vehicles coming to a garage to have a repair carried out, the massive quantity of data available highlights some distinct differences in different types of vehicles. Data on the vehicle and the parts fitted are integrated and checked in preparation for analysis. Data dimensions include make and model of vehicle, type of repair, age of vehicle, date of visit to garage, mileage at time of repair and parts fitted. The data were analyzed by constructing empirical cumulative distribution curves for mileage for specific makes and models of vehicle and type of repair. Figure 2 compares the curves of mileage in vehicles coming to the garage for brake disk replacement for 3 different popular saloon cars. The plot shows that cars of type A go to the garage with much lower mileage than for either of B or C. In this data set 50% of cars of type A have mileage 65,000 or less whereas for cars of types B and C the percentage is nearer 25%. This implies that cars of type A have brake disc replacement earlier and are in this sense a less desirable car.

>> Insert here Figure 2>>

Figure 2 Brake Disc replacement for 3 popular saloon cars.

The outcome of the Data Science analysis is an empirical tool that can be updated as more data arise and can be used by the garage to determine necessary stock levels, likely co-morbidity when vehicles come in for a particular repair and a service to customers when considering buying a used vehicle.
Secondly, return rates for specific items bought from a catalogue are shown to differ widely; those items that exceed control limits in a funnel plot of return rates merit further investigation which may include checking how the parts are presented in the catalogue.

Thirdly, only the original manufacturer knows exactly which original equipment was fitted to a specific vehicle and generally their parts are more expensive than the generic copies which are made by other companies. There is always some uncertainty when buyers are choosing between alternative products and this opens the way to some business opportunities using data analytics. Suppliers of a particular spare part for a vehicle were interested to know if their part would fit vehicles other than the one for which it was designed. A search through the database of part dimensions and a matching process led to identification of several other vehicles that could also use the part (24).

This case study looks at adding value to administrative and operational data from a range of company sources and demonstrates where improvements can be made in 3 areas: in the service offering by garages; in catalogue clarity and usability; and in reach for suppliers.

Shipping sector

The shipping industry is the nucleus of global trade and is highly sensitive to fuel prices. Fuel costs represent over 50% of the total operating costs of a vessel (25) and contribute a significant portion of the total transportation cost of cargo. Fluctuations in the price of crude oil and stricter environmental regulations on the emission of noxious and greenhouse gases are influential factors in the operation of the shipping industry. Data Science has been used to develop new products offered by a small to medium enterprise (SME) dealing with shipping performance data.

Sensors attached to ship's engines record fuel consumption which is displayed in real time on the ship's bridge. As well as providing a check on fuel theft and gross errors in engine function, this data can be used for many other purposes. The business motivation of this case study is to provide a decision support tool to aid the scheduling of ferry services.

Fuel consumption is related to the ship's speed over ground. The Company operational data is especially useful when enhanced with open data on weather and tides. The components of weather that most affect fuel consumption have been shown to be head wind and cross wind. Tidal data is important in certain shipping routes and in this case study, the tides around the UK are quantified using algorithms and data on freely available websites.

Company data on fuel consumption over the whole journey of a ferry carrying out a daily return service on a particular route is integrated with weather and tide data after adjusting them for location and time granulation issues. The data are analyzed using multivariate statistical analysis to produce a predictive model for fuel consumption based on known tides and predicted weather.

>> Insert here Figure 3 >>

Figure 3 SPC chart for fuel consumption. UCL and LCL are upper and lower control limits. Lines for Predicted and actual fuel consumption (FC) are shown.

Figure 3 shows a statistical process control chart constructed using the residual fuel consumption after fitting a regression model. It can be seen that during the time shown in the control chart, the outward journey generally requires less fuel than the return journey due to the tide. Journeys that have fuel consumption within the upper and lower control limits allowing for the specific weather and tide can be considered as satisfactory. Journeys exceeding either the upper or lower control limit merit further investigation as to possible causes for the fuel used. In this case, all journeys were within the control limits. Using control limits avoids the shipping company unnecessarily spending time checking when there is no real change in the performance of the vessel.

Data Science enables the shipping company to understand the performance of its ships and the variation in fuel consumption costs and corresponding emissions due to weather and tides. Where sufficient flexibility exists, journeys can be scheduled to minimize cost. In this way the analysis provides a valuable management decision making tool based on data already collected and available.

In another application of Data Science in the SME, the most economical speed for the vessel to travel is estimated. Ships are subject to dock trials while the ship is being constructed, builder's trials when the ship is completed and sea trials when the ship starts in service. These trials indicate the economical speed for travel. However, over time, the ship's performance changes due to fouling, damage caused by minor collisions, etc. Employing statistical thinking to design trials for the ship during service leads to a reassessment of the most economical speed and consequent improvement in the business.

This example looks at adding value to fuel consumption data collected for the purpose of monitoring, and developing new products for the SME thus increasing sales and widening their customer base. The work involves detailed knowledge of the shipping scenario, handling and amalgamation of vast sets of fast moving data and access to open data, statistical analysis and then business know-how to convert the findings into a valuable product for the company to sell. It shows the value of the combination of all 3 aspects of Data Science.

Social housing sector

Social housing accounts for about 50% of rental properties in the UK. It aims to provide affordable accommodation and is controlled by strict governmental regulations. Certain social housing providers have tens of thousands of properties, but still rent collection, repairs and maintenance are usually outsourced to bespoke software providers. The business motivation of this case study is to identify tenants in danger of falling into arrears before their debts become too difficult to manage. These tenants can then be assisted to make their payments and the housing provider can potentially reduce the level of their arrears caseload and account processing time, thus increasing income collection.

In this example, a small to medium enterprise (SME) software house is using Data Science to maximize the insight that can be obtained from their vast reservoir of interesting data on rent balances, property repairs and empty properties, or voids (26). Social housing data is complex and includes information about the property, about the tenant and the payment details during each tenancy. This is confidential data that has to be handled with absolute adherence to data protection laws. There needs to be extensive data amalgamation from multiple tables dealing with property characteristics, tenancy information, ongoing payments and interactions between the tenant and rent officers. Hence, preparing the data for analysis is a major part of the project.

Weekly arrears data is analyzed using time series and machine learning methods including cluster analysis. Figure 4 shows the main pattern of rent balance profiles over different periods of time. Balances are negative when rent has been paid and thus the tenant is in credit, while positive balance means there is a debt. Cluster analysis identified a major cluster of tenancies and multiple exceptional clusters. In Figure 4, the left-hand side main cluster illustrates temporal patterns. The top plot shows that rent arrears follow a downward trend, while the middle plot shows that rent arrears increase throughout the month and the lower plot shows the peak in winter. The plots on the right-hand side of Figure 4 demonstrate typical outliers, with an overall upward trend over the three-year period in the top plot and no clear monthly pattern in the middle plot; the lower plot shows a more variable yearly pattern than the main cluster with a shorter period of being in credit than for the majority.

>> Insert here Figure 4 >>

Figure 4 Cluster analysis of rent balance profiles

Statistical and machine learning analysis have been used to predict arrears, to model the weekly, monthly and yearly patterns in rent balances and to identify clusters of tenants and understand the important features in each cluster. In addition, data visualization has been shown to be a very powerful tool for giving social housing providers insight into their business. This example illustrates a Data Science application in a sector hitherto unfamiliar with data analytics and shows that there are enormous benefits from using vast routinely collected operational data resources for the purpose of day to day running of the business, in a sensitive way.

7. Final Comments and Future View

We would like to mention a representative episode from a New York Times article in 2004 (https://www.nytimes.com/2004/11/14/business/yourmoney/what-walmart-knows-about-customers-habits.html) that is mentioned in Foster Provost F. et al. (27):

> *Hurricane Frances was on its way, barreling across the Caribbean, threatening a direct hit on Florida's Atlantic coast. Residents made for higher ground, but far away, in Bentonville, Ark., executives at Wal-Mart Stores decided that the situation offered a great opportunity for one of their newest data-driven weapons … predictive technology. A week ahead of the storm's landfall, Linda M. Dillman, Wal-Mart's chief information officer, pressed her staff to come up with forecasts based on what had happened when Hurricane Charley struck several weeks earlier. Backed by the trillions of bytes' worth of shopper history that is stored in Wal-Mart's data warehouse, she felt that the company could 'start predicting what's going to happen, instead of waiting for it to happen', as she put it.*

The authors wonder why data driven prediction is useful in this situation: for predicting the need of an increased amount of water bottles? Or for discovering that a particular product was sold out in those days and this fact may be related to the hurricane? Or trying to learn from the data in previous hurricanes if there were particular needs? We may go on with more points at issue, but these facts are not difficult to be predicted or to be explored. It's surely more interesting to look for more sophisticated and more aimed models, mining the huge amount of data stored by Wal-Marts before the hurricane Florence. And so it was: The New York Times article (Hays C.L., What Wal-Mart Knows About Customers' Habits, November 14th, 2004) continues:

> *… the experts mined the data and found that the stores would indeed need certain products —and not just the usual flashlights. "We didn't know in the past that strawberry Pop-Tarts increase in sales, like seven times their normal sales rate, ahead of a hurricane," Ms. Dillman said in a recent interview. "And the pre-hurricane top-selling item was beer."*

We report the anecdote to highlight the main strength of Data Science: the Data-Driven Decision Making i.e. the objective of coming to decisions based on rigorous and accurate analysis of data rather than on intuitive insight alone. It has been shown that data dependency in a company makes it at least 6% more productive. Business needs in-depth analysis and to be provided with a detailed overview of the current situation and of the prospective changes.

Unfortunately, as the change in business models by companies is to become more and more data driven, they come up against the inadequacy of supply of qualified personnel, i.e. Data Scientists: their number is not sufficient in the face of such a fast and radical change in business. For this reason, universities are preparing academic and post-academic courses with different but relevant curricula. This testifies to the good job prospects for Data Scientists and the extent to which the ability to analyze data should become a key tool for all ambitious managers.

In order to ensure good practice, specialists in the different parts of Data Science need to engage with each other and embrace their different ways of thinking. It was realized with Six Sigma that a business focus was vital and this is even more the case with Data Science. The challenges that are tackled and the solutions found must be tailored to issues of strategic importance to the business. Otherwise the Data Science becomes more of an academic exercise and companies downplay the role that Data Science could have in their company thus restricting the opportunities for innovative applications. Data Scientists need to be creative in their thinking, central to the business and constantly mindful of the advantages of having an open and free exchange with specialists in different areas. Good Data Science requires effective communication with other professionals and for all to act together in harmony.

Statisticians need to be confident that what they have to offer is valuable and important. Companies need help to understand this new data centric world and one way is to present case studies which clearly show the benefits of correct analysis and its supremacy over black box solutions. Reports of successful and less successful Data Science projects need to be readily available not only in academic publications but also in the trade press with which business people are familiar.

This requirement is being addressed by a growing literature of case studies and opinions available in the popular press, for example the article by Shan (28), and in specific issues of journals, for example the paper by Giudici (10). Text books are slower to follow suit but there are some excellent publications incorporating statistical techniques in a Data Science context, such as James et al (29) and the more all-encompassing Data Science Handbook (30). Our future view is one of optimism for the continued progress of statistical involvement in Data Science and the increasing importance of Data Science as a modus operandi for companies. The enormous financial benefits on offer ensure that there is adequate funding for research and new ideas and methods are emerging all the time.

References

1. **Coleman, SY.** Data science in Industry 4.0. [book auth.] ECMI. *ECMI conference, Budapest June 2018, in ECMI book subseries of Mathematics in Industry.* s.l. : Springer, 2019.

2. **Ahlemeyer-Stubbe, A and Coleman, SY.** *Monetising data – how to uplift your business, Wiley.* London : Wiley, 2018.

3. ***Naur, Peter. "Concise Survey of Computer Methods". Lund, Sweden: Studentlitteratur. 1974. Retrieved from: http://www.naur.com/Conc.Surv.html.***

*4. **Joseph, Hugh A. Chipman and V. Roshan.** A Conversation with Jeff Wu. Statistical Science. 2016, Vol. 31, 4, pp. 624–636.*

*5. **Wu, C.F.J. "Statistics = Data Science?" 1997.***

*6. **Flach, P.** Machine Learning: The Art and Science of Algorithms that Make Sense of Data. s.l. : Cambridge University Press, 2012.*

*7. **Breiman, L. (2001) Statistical Modelling: The Two Cultures. Statistical Science. 16(3), 199-231.** , p. .*

*8. **Box, G. E. P. (1976). Science and Statistics. Journal of the American Statistical Association, 71: 791–799. doi:10.1080/01621459.1976.10480949.***

*9. **Cleveland, W. S. (2001). Data science: an action plan for expanding the technical areas of the field of statistics. International Statistical Review / Revue Internationale de Statistique, 21–26.***

*10. **Giudici, P.** Financial data science. Statistics & Probability Letters. May 2018, Vol. 36, pp. 160-164.*

*11. **Wired., Chris Andersen (2008) The end of Theory: the Data Deluge makes the Scientific method Obsolete:.***

*12. **Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. From Data Mining to Knowledge Discovery in Databases. Al Magazine, 17(3). 1996. Retrieved from: https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131.***

*13. **Brachman, R., and Anand, T. 1996. The Process of Knowledge Discovery in Databases: A Human-Centered Approach. In Advances in Knowledge Discovery and Data Mining, 37–58, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Menlo Park, Calif.:.***

*14. The Great Expectations of the ImageNet Challenge. **Soft., Science.** s.l. : https://www.scnsoft.com/blog/imagenet-challenge-2017-expectations, 2017.*

*15. **Efron, B and Hastie, T.** Computer Age Statistical Inference: Algorithms, Evidence, and Data Science. s.l. : Cambridge University Press, 2016.*

*16. **Patil, d and Davenport, TH.** Getting control of Big Data. Harvard Business Review. 2012.*

Hgigit

17. Coleman, SY and Kenett, RS (2017) The Information Quality Framework for Evaluating Data Science Programs. Available at SSRN: https://ssrn.com/abstract=2911557.

18. Brown, J. (2018). Leading change: Developing a new Applied Data Science programme. ICOTS10, Kyoto, http://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_1H1.pdf?1531364187.

19. kdnuggets discussion. [Online] 2013. https://www.kdnuggets.com/2013/04/data-science-end-statistics-discussion.html.

20. Coleman, SY, Gob, R, Manco, G, Pievatolo, A, Tort-Martorell, X, Reis, M (2016) How Can SMEs Benefit from Big Data? Challenges and a Path Forward. Journal of Quality and Reliability Engineering Int., http://onlinelibrary.wiley.com/doi/10.1002/qre.2008/full.

21. Coleman, S.Y. Six Sigma – an opportunity for statistics and for statisticians. 2008, Vol. 5, pp. 94-96.

22. Mustafazade, F. (2018). Using social science data to solve a social housing problem. https://blog.esrc.ac.uk/2018/10/19/using-social-science-data-to-solve-a-social-housing-problem/.

23. Smith, W, Coleman, S, Bacardit, J, Coxon, S. Insight from Data Analytics with an Automotive Aftermarket SME. Quality and Reliability Engineering International. 2019, pp. 1-12.

24. Smith, W, Coleman, S, Bacardit, J. and Coxon, S. (2018) How data can change the automotive aftermarket. Focus, p30-32, October, www.ciltuk.org.uk.

25. Zaman, I, Pazouki, K, Norman, R, Younessi, S, Coleman, SY. Development of automatic mode detection system by implementing the statistical analysis of ship data to monitor the performance. Int. J. Maritime Engineering, RINA Trans A3. 2017, Vol. 159, pp. 225-35.

26. Mustafazade, F, Coleman S, Bacardit, J (2018). Application of machine learning for decision support in social housing. Statistics and Data Science - new Developments for Business and Industrial Applications conference, Turin, http://www.sds2018.polito.it.

27. Provost, F., Fawcett, T. Data Science for Business - What you need to know about Data Mining and Data-Analytic Thinking. s.l. : O'Reilly Media, USA., 2013.

28. Shan, Carl. What-are-good-examples-of-using-data-science-for-development-and-or-social-good. quora.com. [Online] 2014. [Cited: 2 August 2019.] https://www.quora.com/What-are-good-examples-of-using-data-science-for-development-and-or-social-good.

29. James, G, et al. The Introduction to Statistical Learning with Applications in R. s.l. : Springer, 2017.

30. Cady, Field. The Data Science Handbook, Wiley. s.l. : Wiley, 2017. ISBN: 978-1-119-09294-0.